


## Appendix 1. BLAST alignments

### A. hepsin vs spinosin:

 **Blast 2 Sequences results**

PubMed Entrez BLAST OMM Taxonomy Structure

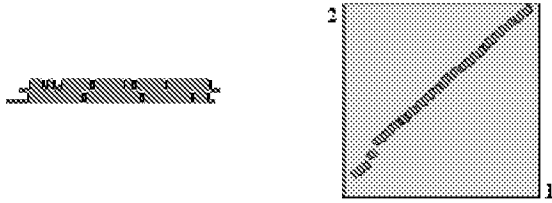
BLAST 2 SEQUENCES RESULTS VERSION BLASTP 1.2.17 [Aug-26-2007]

Matrix: **BLOSUM62** gap open: **11** gap extension: **1**  
x\_dropoff: **0** expect: **10.0000** wordsize: **3** Filter: ☐ View option: **Standard**  
Masking character option: **X for protein, n for nucleotide** Masking color option: **Black**  
☐ Show CDS translation **Align**


---

Sequence 1: [gi123057](#) Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]  
Length = 417 (1 .. 417)

Sequence 2: [gi1224891](#) Spinosin [Homo sapiens]  
Length = 457 (1 .. 457)



NOTE: Bitscore and expect value are calculated based on the size of the nr database.



Score = 221 bits (588%), Expect = 1e-55  
Identities = 140/412 (34%), Positives = 200/412 (48%), Gaps = 51/412 (12%)

Query 24 GTLLLLTAIGRAASRAIVAVLLRSQGEPLYPV-----QVSSADARLMVFDEK--- 69  
G L LL G SN +V L + +S+ +S+ +A L KI  
Sbjct 63 GALLLLAGAGVGSMLLVLYLCPANSGSISQTLQDEEITLSCSEASAEETALLPALPMTVSE 112

Query 70 -----EGTWELLCSSRSNRVRGLSCGFESGFLRALTHSELDVWTAGANGT 114  
+ W L+C + + C +G LR H +++ N +  
Sbjct 113 RINSEDFLLAQVEDQPSWLLVCHEGNSPALGLQIOWSLGHLRLTHKKQFVNLTHIKLNE 172

Query 115 SGFFQVDEGRLEPHTQRLLEVI--SVDDCPGRFLAICQSGGRKLEPVDREIVGGRTSLG 172  
F + F LE +C G+ ++ C +CG R L RIVGG+ + G  
Sbjct 172 QETFAQLS----PALGQFLZEAMQPPNNCTSGQVVSLSCTSGGAPPL-ASRIVGGQSVAFG 227

Query 173 RNFQCVSLRYDGAHLQCGSLLSGDWLTAARCFEE-RNENLGRMENVFASVAQAS--SHG 229  
RNPQ2 S+ H CGGS+L+ NV+TAARD R LS WRV AG V+ ++ SH

```

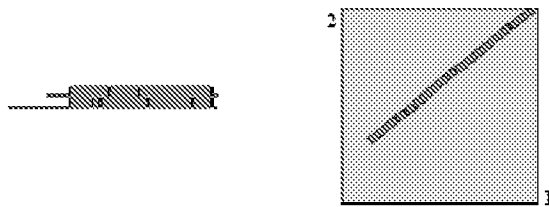
Subject 228 KKEWQAEVALGERHTCGGSVLAPRWVIAAHCKHSEFLELLESEWRVHESLVSHEAVRSHQ 227
Query 230 LQLGVQAVVVHGGYLSPEDNSZENSNDIALVHLSSFLFITEYIQSVOLPAAQQALVDGK 229
      L V+ ++ R Y      +++ + D+AL+ L + L ++ + VILPA Q   G
Subject 238 GAL-VERIIFHELI-----SAQNHEDVDVALLALQIALNFSDIVGANVLPAGEQHEFKGS 245
Query 236 ICTVTCWENTQ-YYQQAGSLQBARVFIISNDVYNGADFYGNQINPKNFORGVFEKGIDA 248
      C V+GSG+T   +   + +LQ+ VP+ S +GN + Y   + P+N CAGI +G DA
Subject 241 RQVPSGNGHTHEESHYSSERMLQETVVPLFSTQLONSSGVYSGALTFEHLCAQYLDGRADA 400
Query 249 CQGDSSGSEFFVCEDSISRIFFNRLOGIVSWGTGCLACNFGVYTHVSDPREWT 400
      CQGDSSGSEF VD D      NPL G+VSWG CA      DEVY NV++F +WI
Subject 401 CQGDSSGSEFLVCPDG----DTWRLLGVWSEKACAZEPNHPGVYAKVAEFLSWI 445

```

## B. hepsin vs TMPRSS2:

**Sequence 1:** [gi|123037|Serine protease hepsin \(Transmembrane protease, serine 1\) \[Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain\]](#)  
Length = 417 (1 .. 417)

**Sequence 2:** [gi|14503498|transmembrane protease, serine 2 \[Homo sapiens\]](#)  
[hg25061724hg.A4D37117.1:AF123453\\_1 transmembrane serine protease 2 \[Homo sapiens\]](#)  
Length = 492 (1 .. 492)



NOTE: Bitscore and expect value are calculated based on the size of the nr database.



Score = 210 bits (53%), Expect = 3e-62  
Identities = 117/346 (34%), Positives = 174/346 (50%), Gaps = 17/346 (4%)

```

Query   68  LNVPEKTEPTWELLCSRRNRVRVGLSCCEEMGLRALTHSELVPTAGKSTGPPFVDE 122
      L ++      -K +G   N      +D +H3+      S+ Y  +G+
Sbjct  159  LQMYSSQRKSNWHPVQGDWZNYGSAACRDNGYKNNFYSSQGVNDSGSTSFNKLN-TSA 216

Query   123  GRLEPKTQRLLELVISWEDCFRGRFLAAICQDCG--REKLEVDSEIVGGRDTSLGRNPKQVEL 186
      G +   ++L   +  C   ++ C  CF      REWGS   G NFRQVEL
Sbjct  217  GRVDIYHMLYHSDA--CQSKAVVSLRLCIQGVNLSGSRQSRIVGSEELPGRNFWQVEL 270

Query   181  SYDGAHLGGGSLLSGKVVLTAAHCFPEPNERVLSRWVUFAGAVAGASP-HGLQLSVQRAVY 239
      H+CGGS+++ +N++DALHC +      W PEG + Q+  +G   V+ V+
Sbjct  274  HTQNVHVCQGSIIPTENIVTAAKOVKSLRNPNHNTAFAGILRQSFNFGAGYQVEKVIS 333

Query   240  HGGHLSFRDPNSENENIDIALVHLSSFLPLETYIQPVLPAAAGQALVDGNICTVTCWQNT 299
      H Y      +S+ +NIAL+ L  PL  + ++PVCLP  G  L  ++C ++GWS T
Sbjct  334  HPNV-----DSKTKNNDIALMHLQKPLTFNDLVKGVCLSPGRMLQPEQLQWISGKGT 397

Query   300  QYVEQQAGWLGQAPUPLIINDVCKGADFYGNQINRMPGAGYEPGQIDACQGDSSGSPFW 359
      + G+ + VL  A+V +I   CN   Y N I P N CAG+ +G +D+CGGSGSP V
Sbjct  355  EENQNTSEVLMNAKVVLIETQRCHSRVYVDNLITPANCAGSLQGNVDSGQGDGPPPLV- 446

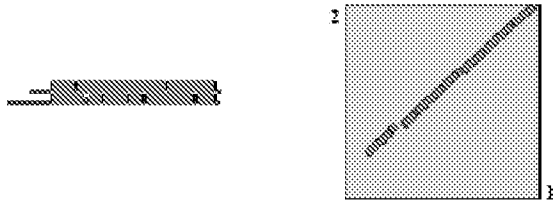
Query   365  EDSISRTFRNRLCGIVNKGTSCLAGCPGVYTKVSDFRKNIFQAIK 405
      S+   W L G  SWG+GCA A +EGVV  V  P +NI++ ++
Sbjct  447  ---TEKNINWNLIDETSWGSGCAKAYRQGVYCNFVPTDNIYRQMR 509

```

### C. hepsin vs TMPRSS3:

Sequence 1: [gi|273057|Serine protease hepsin \(Transmembrane protease, serine 1\) \[Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain\]](#)  
Length = 417 (1 .. 417)

Sequence 2: [gi371871821MPRSS3 \[Homo sapiens\]](#)  
Length = 432 (1 .. 432)



NOTE: Bitscore and expect value are calculated based on the size of the nr database.



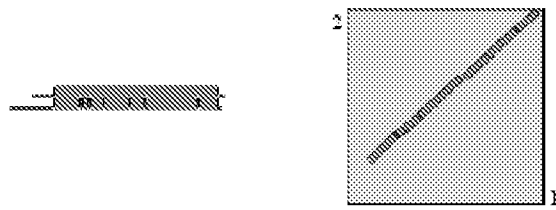
Score = 221 bits (88%), Expect = 9e-56  
 Identical = 129/360 (36%), Positives = 172/360 (48%), Gaps = 34/360 (9%)

Query	80	FLYFWVSSADARLNGVDFTEATWRLLOSSKARVAGLSSTEEMFPLALT---HSELOV	106
		P V++S + L V D G N C +A +C +MS+ SA+ +LW	
Subject	106	FAFWNLKSDSTLQVLSDATGNVFCDFDFTLALATACQPMQYSRAVEIGPQQLLV	155
Query	107	RTAGAKRSTSGFFVDEGRLESHFQRLELIVSVQDQPMGRFLAALCQDQERRKLEVDRIWQ	168
		M +Q L S C G ++ C CG+ L E+VQ	
Subject	160	VEITEN-----SRLPMRNSGGCLGSSIVLNLALDGN-SLTFPRVWQ	203
Query	167	SDILSPNQWQVSLAYDCAHLCCGLSLGGNVLTAARCPFFRNFTSLKRWVATAGAVAGS	216
		+E+ NQWQS++YD H+CGS+L SVLTAAHCP + V +W+V AG+ E	
Subject	204	ZEASVDSNPWQVSIQVDEKQHVTCGSLDPLHNVLTAAHCFKAKHTDFTN-WKVRACGOWLGS	262
Query	227	PHGLQSLVCAVYVHSGFLPFEDNSEENSDIALVLHSSFLSLTIVQFVCLSAAGQALV	268
		LQ+ ++G+ F NDIAL+ L PL + ++P+CLP + L	
Subject	263	PSRLAVAKIITIEEFNWFD-----KNDIALHKLQFLTFSGVTRPCLTFDEBLT	314
Query	267	DNKICTVDEGHT-QHYGQDAGVLQEARVETISNEVCGMGAFYQWIKNFKFCAGYPEG	345
		+GRG T Q G++ +L +A V +I + CN D Y ++ KM CAG PEGG	
Subject	318	PATPLWIIIGSPTQNGGCMRSDILLQAEWQVLDSTPNADDAVQGEVIENHNCAGIPGG	374
Query	346	IDACQDRESGFFVCHDSISRTFARKLLOGINWSTSCALAAQKGVYTVKVDFFENIFQAIK	408
		+D CQGDGSGG + + + +N + GIVNSG CGG EGYTYS + W+ K	
Subject	375	NDTCQDGGGGLHKT-----SDQSHVTVKSGVTCGCGGFPEDSVTVNVAIVNLWIKWVK	421

## D. hepsin vs TMPRSS4:

Sequence 1: [gi123057:Serine protease hepsin \(Transmembrane protease, serine 1\) \[Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain\]](#)  
Length = 417 (1 .. 417)

Sequence 2: [gi15451940:transmembrane protease, serine 4 isoform 1 \[Homo sapiens\]](#)  
[gi13633980:sp.Q00784.TMPRSS4\\_HUMAN Transmembrane protease, serine 4 \(Membrane-type serine protease 2\) \(MT-SP2\) >gi18347149:gi1AAF74436.1:AAF179224\\_1 transmembrane serine protease 3 \[Homo sapiens\] >gi13079795:gi1AAH31703\\_1 Transmembrane protease, serine 4 \[Homo sapiens\]](#)  
[gi147077743:cds:BAD18749\\_1:unannoted protein product \[Homo sapiens\]](#)  
Length = 437 (1 .. 437)



NOTE: Bitscore and expect value are calculated based on the size of the nr database.

Score = 215 bits (558), Expect = 4e-55  
Identities = 126/357 (35%), Positives = 180/357 (50%), Gaps = 23/357 (6%)

```

Query    50  PLYPWQVSSADARLNVFDMTEGTWALLCSRSNARVAGLSCBEMGSLRALTHSELSWETA 109
          S++E++LVDGK+C+EA+CG+I+++
Sbjct    106  PAVAVLSNDRSITLQVLSATGNWPSACFDMTEALAEATCRMGYSKPTTPAVEI--- 156

Query    118  GANGTSGFTCVSEGLPHTQRLLENISVDCDPRGAFLLAICDCCGRNLEVDRIVGGRTI 165
          G++E++QLS+CG++CQG+LRVGG+
Sbjct    157  GPDQDLDTVEITE---NRQLRMNRSSGPGLSGLVSLACLACGK-SLNTPRVUGGEZA 211

Query    170  SLGSHDHWQLRYDCAKLOGGSLSCDWVLTAAHOFPERNRVLSRWVFAQAVAGQASPHG 228
          S+ NWQVSS++VDH+CGGS+LWVLTAAHCF+V+N+VAG+S
Sbjct    212  SVDSWNRQVNEIQVDMGHVCGSRILDPHWVLTAAHCTAKHTAVPN-NWVRAGSDMLGSEPS 270

Query    230  LQLGVLAVVTHGGYLPFRDFNSEENNDIALVHLSSFLTEYIQPVCLPAGQALVDGN 288
          L++P+NDIAL+LPL++P+CLP+L
Sbjct    271  LAVAKIIIIIFRHWV-----KNDIALNHLQFLIFSCTVARIQLSFFDEELTPAT 321

Query    290  ICTVTEGKNT-QYVQQAQVLQEARVPIISNDVCGAGSYGNQINENKPCAGYPEGGIDA 348
          +GRG+TQG++L+AV+I+CN+D+Y++NMGRGPEGG+D
Sbjct    313  PLNIIIGQFTKQNGGGRSDILLQASVQVLDSTRGNADDAVQGVTEKQKAGIPEGGVDT 361

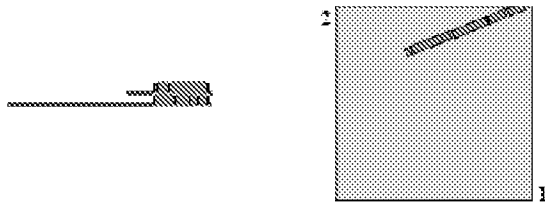
Query    345  CQGSQGGFFVCEDSISRTFANKLGGIVERSTGCLAQKPGVYTKVSDFRKIFQAIN 405
          CQGSQGGF++ ++N+GIWSNGGC+GVYTKV++N+Y
Sbjct    383  CQGSQGGFLNKG-----SDQHWVTVIVSNGVCGGSGSTFGVYTKVSAKLNVINWVK 434

```

## E. hepsin vs enteropeptidase:

Sequence 1: [gi1230377](#) Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]  
Length = 417 (1 .. 417)

Sequence 2: [gi5690001](#) enteropeptidase (Homo sapiens)  
Length = 1019 (1 .. 1019)



NOTE: Bitscore and expect value are calculated based on the size of the nr database.

Score = 111 bits (53%), Expect = 1e-52  
Identities = 138/256 (42%), Positives = 145/256 (52%), Gaps = 17/256 (6%)

```

Query 151 QDCGRRLKND---PIVOCEDTSLGRNFKQVSLRUDGANLQGGSLSGENVLTAAHCFEE 207
+ CG++ D +IVQG + Q WPN V L Y G LQG SL+S DW+++ARHC
Sbjct 770 KSCGKKLAACDITIEIVGSSNAKEGANSFVVGGLYGGRLLOGASIVSSDWLVSAAHCVIG 828

Query 205 ENRNLSEKRVYTAGAVAGA---SPHQLQLGVQAUVVHGGYLFREDPHEENNDIALVHLS 264
EN S+W G ++ SP + + +V + Y N NDIA++HL
Sbjct 890 ENLEPSKNTAILGLRKEENLTPSPQTVERLIDEIVINSHY-----NDRKNDIAMHLE 948

Query 268 SPPLSLTPTIQQPVLPAAQALVDGNIQIVTQWQNTQYVGGQAGVLQEARVPIISNDVQNG 324
+ I+YIQP+CLP Q G+ C++ QNT Y G A +LQEA VP++SN+ C
Sbjct 984 FFWNTDYTIQPIQLFEENQVFFFGRRNSIAGWGTVVYQGTTANILQEAADVLLSNEROQ- 1042

Query 328 ADFVGNQIKPKMFCAGYEEGGIDACQGGDSGGFFVCELSISRTPSNPLOGIVSWGTGCALA 384
I M CAGY EGGID+QGGDSGGF +D+++ RN L G+ S+G CAL
Sbjct 1043 QQMPYNYITENMKICGVYEEGGIDSCQGGDSGGFLAQEN---NKNFLAGVTETGTCWALP 1098

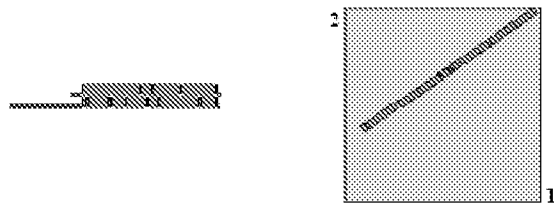
Query 388 QKPGVYTKVSDPRENI 400
+DGVY +VS F ENI
Sbjct 1099 NRKGVYAKVSDPRENI 1014

```

## F. hepsin vs MSPL:

Sequence 1: [gi123057](#) Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]  
Length = 417 (1 .. 417)

Sequence 2: [gi116356363](#) transmembrane protease, serine 13 [Homo sapiens]  
Length = 567 (1 .. 567)



NOTE: Bitscore and expect value are calculated based on the size of the nr database.

```

Score = 221 bits (588), Expect = 1e-53
Identities = 134/372 (36%), Positives = 196/372 (52%), Gaps = 18/372 (7%)

Query   38  IVAVLLRSDEQEELYPVQVSSADARLAVFVKTEGTWELLCSSPSNARVAGLSCEHKEFLPA  58
      +V  L+EE+ L V+      + L ++ + W +DS N + +G+++GF A
Sbjct  214  VVDCNLSDE--LGVRFVNDKELLKIYSGSSHQWLPICSSNWNDSYSEKTOQQLGFESA  271

Query   59  LTSSSELVRTAGAGTSGFPVDEGRLESTQRLLEVISVDCDPSGPELARIQDDGRNKL  159
      +E+ R      F      L + + E + +CP E+++ C CG R +
Sbjct  272  HRTTEVAHRD-----FANSEFILRYNSTIQSELHREBCPSQRYISLQCSHGLEAN  322

Query  159  PVERIVGQPDTSLGKMPKGVSLKEDARHLCQGSLLSGIKWLTAAHC-FPERNVLGRNV  217
      RINGG S +NPKQVEL + H+CGG+L+ KVLTAHC P R +VL W+V
Sbjct  323  -TSPVVGALASDSKWPQVSLHFGITHICGGTLIDAQWLTAAHCOFFVTREKVLGKNV  381

Query  218  FAGAVACASPHSL--QLGVQANVYHGGVLPFRDSEENSDIALNHLSSPLFTEVIGS  275
      +AG ++ H L      + ++ + Y      EE+ DIAL+ LS SL L+ +I P
Sbjct  382  YAGT---SNLRLQPEAASIAETIENSIV-----TDERDQYDIALNRLSNPLTSAHNP  432

Query  278  VOLPARAQALVDGNIOTVDSKQNTQYYQDQAG-VLQEARVFIENQVONCARPYGNQIKP  334
      CLP GQ + C +EG+G P+      + L+E +V +I CN Y + + P
Sbjct  433  ACLPMHGGITSLNITQNTGPGWTFRETDDKTSFPLKEVYVHLIDEPYONDLYVDGYLIF  491

Query  386  RMFCAGYHEGGIDAQGGSGGFFVUEDSISRTPRNRLOGIVNGTGCALAQKPGVYTKVS  444
      +M DAP GQ D+CGDSGGP VCE + PE L G+ SNGTGG KPGVYTKV+
Sbjct  492  RMFCAGNLRGGDRSCGGSGGSLVCE---QNNRWYLAGVTSWGTGCGQENKPGVYTKVT  548

Query  395  DEPEKIFQAIKI  456
      + WI+ +++
Sbjct  549  EVLFWIYSNME3  600

```